



The effects of survey mode on self-reported psychological functioning: Measurement invariance and latent mean comparison across face-to-face and web modes

Gaja Zager Kocjan¹ · Darja Lavtar² · Gregor Sočan¹

Accepted: 29 April 2022 / Published online: 26 May 2022
© The Psychonomic Society, Inc. 2022

Abstract

Despite the increasing popularity of mixed-mode approaches to data collection, studies examining measurement equivalence across different survey modes in representative samples of the general population, particularly focusing on measures of socially sensitive psychological constructs, are sparse. In this study, we used data from a large representative sample of the Slovenian population ($N = 9,900$) collected as part of the third wave of the European Health Interview Survey (EHIS) to examine mode-inherent effects (i.e., measurement effects that cannot be neutralized by clever survey design) of the traditional interviewer-mediated face-to-face mode and the increasingly popular self-administered web mode on three measures of psychological functioning, namely the Patient Health Questionnaire Depression Scale (PHQ-8), the Satisfaction with Life Scale (SWLS), and the Mental Health Continuum – Short Form (MHC-SF). After controlling for self-selection propensities, our results showed strict measurement invariance for all three scales across the two survey modes, but statistically significant and notable differences in latent means, suggesting that individuals who respond face-to-face systematically report better psychological functioning than individuals who respond over the web. These results suggest significant mode inherent effects that cannot be attributed to measurement non-invariance of the scales between face-to-face and web survey modes, but most likely to social desirability bias in responses achieved in the presence of an interviewer. Administration mode effects must be considered when interpreting and comparing results obtained through different survey modes, particularly interviewer-mediated versus self-administered modes, especially when using measures of culturally desirable traits and behaviors, such as mental health and well-being.

Keywords Administration mode effects · Measurement invariance · Social desirability bias · Psychological functioning

Introduction

Many recent population surveys combine different modes of data collection due to several benefits, including lower costs and increased response rates. However, the administration mode may represent a source of measurement bias (De Leeuw et al., 2019; Hox et al., 2015). Therefore, it is essential that before comparing scale scores between survey

modes and possibly merging them into a single dataset, measurement equivalence across modes is established, and potential systematic differences between modes are examined. Nevertheless, studies on cross-mode measurement equivalence in representative samples of the general population, particularly those involving a combination of more traditional face-to-face mode and increasingly popular web mode, remain scarce, and even rarer are those focusing on measures of socially sensitive constructs (e.g., Zhang et al., 2017). To advance literature in this field, the present study uses representative data from the third wave of the European Health Interview Survey (EHIS 3) in Slovenia to examine measurement invariance and compare latent means for three scales measuring positive and negative indicators of psychological functioning across face-to-face and web modes of data collection.

✉ Gaja Zager Kocjan
gaja.zagerkocjan@ff.uni-lj.si

¹ Department of Psychology, University of Ljubljana, Ljubljana, Slovenia

² National Institute of Public Health, Ljubljana, Slovenia

Mixed-mode data collection and mode effects on measurement

With the development and increasingly widespread use of computers and the Internet, the web emerged as a prevailing medium for data collection. Self-administration, lower costs, fast and geographically distributed data collection, and timely data are some of the convenient features that make the web mode a promising alternative to traditional methods of data collection, such as face-to-face interviews, telephone interviews, and mail surveys (Callegaro et al., 2015). In practice, the web mode can be combined with other methods of data collection in a mixed-mode survey design. Mixed-mode data collection has become a common practice in modern survey designs as it overcomes the weaknesses of single-mode surveys and minimizes total survey error by improving coverage, increasing response rates, and reducing measurement error (De Leeuw, 2018; De Leeuw et al., 2019; Dillman et al., 2014; Tourangeau, 2017; Tourangeau et al., 2013). Sometimes surveys employ a sequential mixed-mode design. For instance, after an invitation to complete the survey over the web, non-respondents may be contacted by an interviewer for face-to-face data collection (De Leeuw et al., 2019).

In mixed-mode surveys, data collected using different modes are usually merged into a single data set, and sometimes researchers are interested in comparing different modes (e.g., Kreuter et al., 2009). However, survey mode may represent a specific source of measurement bias. Literature suggests that different survey modes may differently affect an individual's response process leading to systematic differences in responses (e.g., Tourangeau, 2017; Tourangeau et al., 2013).

Generally, mode effects can be divided into desired selection effects and undesired measurement effects. Selection effects are related to differences in sample composition across modes since each mode brings in different respondents (e.g., respondents to web surveys are generally younger). Therefore, selection effects need to be controlled for before examining measurement effects (e.g., self-administered modes such as web also elicit less socially desirable responses). Measurement effects result in different errors of observation across modes threatening the validity of cross-mode comparisons. Measurement effects that cannot be offset by clever survey design are mode-inherent effects (De Leeuw, 2018; De Leeuw et al., 2019; Hox et al., 2015, 2017) that are due to inherent characteristics of a survey mode (i.e., the information transmission medium, the presentation and response channels, the involvement of the interviewer, the closeness of interaction between an interviewer and a respondent, and

the use of computer technology in data collection), which determine the underlying principles of communication and transmission between a respondent and the survey questionnaire (Berzelak & Vehovar, 2018). For example, different presentation methods (e.g., visual in web mode and aural, or a combination of visual and aural, in face-to-face mode) may elicit a different understanding of the items and/or response categories leading to different answers (e.g., Christian et al., 2007). Also, the mediation of an interviewer typical for face-to-face and telephone modes, compared to self-completion modes, may evoke more socially desirable responses due to higher consideration of social norms in the context of social interaction (Dillman et al., 2014; Kreuter et al., 2009; Tourangeau et al., 2013).

According to literature (Hox et al., 2015; see also de Leeuw, 1992; Hox et al., 2017; Jäckle et al., 2010), two types of mode-inherent effects can be distinguished. The first type only shifts the location of the response distributions affecting estimates of means but not multivariate estimates (i.e., correlations, covariances). The second type affects the question-answer process. As a result, the questions may be understood and answered differently depending on the survey mode affecting multivariate estimates and potentially leading to measurement non-invariance. In this study, we examined both types of mode inherent effects while controlling for the previously mentioned selection effects.

Measurement and structural invariance in mixed-mode surveys

Comparability of the measured constructs across the administration modes is a necessary precondition for valid and meaningful cross-mode comparisons. Measurement invariance indicates that the construct is psychometrically equivalent and thus measured in the same way in different groups (such as sex, age, education, and cultural groups), time points in longitudinal or intervention studies, and different survey administration modes (Millsap, 2011). In mixed-mode surveys, individuals with the same position on the latent construct (e.g., opinion, attitude, trait) should have the same values on the observed measure of that construct regardless of the survey mode. When measurement bias is introduced, the measure of the construct functions differently depending on the administration mode; thus, merging data or comparing data across modes is precluded. The observed differences may stem either from true differences or from differences in the characteristics of the measurement instrument (i.e., measurement non-equivalence). On the other hand, the observed similarity may result either from true similarity or from differences in the measurement that concealed the true differences. Previous studies on cross-mode measurement invariance vary substantially with regard to modes and measures investigated, not to

mention sampling designs and data collection procedures. Of the studies with representative samples of the general population, few investigated measurement invariance across the web and face-to-face modes; they included measures of constructs related to kinship (Hox et al., 2015), attitudes (Klausch et al., 2013; Revilla, 2013), and aspects of mental health (Zhang et al., 2017). Of particular interest is the study by Zhang et al. (2017), who investigated measurement equivalence of 13 scales measuring positive (e.g., subjective happiness, satisfaction with life) and negative aspects (e.g., pessimism, anxiety) of mental health across the self-administered web and paper-and-pencil modes, and the interviewer-mediated telephone and face-to-face modes. At least partial scalar measurement invariance was established for all the 13 scales across the four administration modes. However, for most of the scales, significant mean differences were found between the interviewer-mediated modes and self-administered modes, suggesting more socially desirable responses in the presence of an interviewer, even when selection effects were controlled.

The latter finding is common in studies examining mode effects on measurement (Berzelak & Vehovar, 2018; Tourangeau & Yan, 2007). It indicates a systematic social desirability bias across all items leading to structural non-invariance with significant mean differences between the administration modes (e.g., Byrne et al., 1989). Such biases typically occur when respondents portray themselves more favorably on measures of socially sensitive constructs. This happens because self-report items contain not only a descriptive/content aspect (e.g., depressed vs. not depressed; happy vs. unhappy), but also an evaluative aspect (e.g., good vs. bad; Bäckström & Björklund, 2014). While social desirability is sometimes described as a stable individual trait that moderates an individual's response to self-report items, it may also be a mediating variable influenced by contextual factors that affect the degree of desirability in self-ratings (Bäckström et al., 2012). Various studies have shown that the presence of an interviewer is one such contextual factor that induces bias due to social desirability, while on the other hand, self-administration curtails such bias (Tourangeau et al., 2013). Self-report measures of attitudes, behaviors, and affective dispositions are particularly affected by the measurement format, with a greater tendency to provide desirable answers in modes involving an interviewer (for an overview, see De Leeuw & Hox, 2011; Tourangeau et al., 2013).

Indicators of psychological functioning in the present study

In this study, we use data from a representative sample from the third wave of the European Health Interview Survey (EHIS 3) conducted in Slovenia in 2019 using a mixed-mode

design. EHIS measures the health status and health determinants of EU citizens, as well as the use of health services and barriers to accessing them. Here, we focus on the cross-mode measurement invariance of three scales used in the Slovenian national EHIS survey that capture different aspects of individuals' psychological functioning and provide data on the mental status of the population. By combining these scales, which are described in detail below, mental illness and mental health were captured simultaneously, overcoming the long-held simplistic view of mental health as the absence of mental illness and disability (Suldo & Shaffer, 2008). Although related, mental health and mental illness represent distinct dimensions, one indicating the presence versus absence of mental health and the other the presence versus absence of psychopathology (Keyes, 2005, 2007; Lamers et al., 2011; Westerhof & Keyes, 2010). Our findings reported in this article show that conclusions about mental health and mental illness in the population may be flawed if we do not account for structural non-invariance between survey modes, which is most likely caused by different levels of socially desirable responding. Indeed, a systematic review (Perinelli & Gremigni, 2016) of studies that have quantitatively addressed social desirability in clinical psychology has shown that social desirability is related to self-reports of mental symptoms, quality of life, and well-being, suggesting a strong evaluative aspect in items measuring these constructs (Bäckström & Björklund, 2014). It has to be noted, though, that most studies included in the review used different scales than those used in the present study.

In the EHIS 3 conducted in 2019 in Slovenia, the illness spectrum of individuals' psychological functioning was assessed using the eight-item version of the Patient Health Questionnaire (PHQ-8; Kroenke & Spitzer, 2002; Kroenke et al., 2009, 2010), which measures current depression and its severity in the clinical and general population. The PHQ-8 is a short form of the widely used PHQ-9 (Spitzer et al., 1999), with nine items corresponding to depression diagnostic criteria according to the DSM-5 (American Psychiatric Association, 2013). In the PHQ-8, the item on suicidal or self-injurious thoughts was omitted because studies have shown that the deletion of this question has little effect on the total score and the two versions of the PHQ have practically equivalent psychometric properties (Kroenke & Spitzer, 2002; Kroenke et al., 2009, 2010). Most previous studies with samples from the general population or from adolescents and college students of different ethnicities supported a unidimensional structure of the PHQ-8 or PHQ-9, consistent with the original model proposed by the scale's authors, and its measurement invariance across various demographic groups (e.g., Baas et al., 2011; Galenkamp et al., 2017; Keum et al., 2018; López-Torres et al., 2019). Although some studies reported an adequate fit of a two-factor model comprising cognitive/affective and somatic

symptoms (for a review, see Lamela et al., 2020), factor structures were different between these studies, and the two factors were often highly correlated ($r > .90$), so several authors advocated for a single-factor model (e.g., Boothroyd et al., 2019; Keum et al., 2018; Pagán-Torres et al., 2020). Measurement invariance of the single-factor model of the PHQ-9 has also been established across telephone and face-to-face modes in a large British patient sample (Ryan et al., 2013).

The positive aspects of human psychological functioning were captured by the Satisfaction with Life Scale (SWLS; Diener et al., 1985) and the short form of the Mental Health Continuum (MHC-SF; Keyes, 2002). The two scales were developed within the positive psychology movement and its broad empirical investigation of well-being constituents and mental health in general (Seligman & Csikszentmihalyi, 2000). Satisfaction with life is often measured in broad population surveys, and it is of great interest in cross-national comparison research (e.g., OECD, 2020; Veenhoven, 2020). However, population surveys often face the problem of length restrictions; thus, single-item measures are commonly used for which psychometric properties are difficult to test. In this respect, the SWLS with five items represents a significant improvement over one-item scales. The SWLS measures a cognitive appraisal of one's life as a whole and thus captures hedonic well-being that, along with a cognitive component, also includes an affective component of positive and negative emotions (Ryan & Deci, 2001). Several previous studies have confirmed the clear single-factor structure of this scale and investigated measurement invariance across various demographic groups (e.g., Emerson et al., 2017; Hinz et al., 2018; Jang et al., 2017; Jovanović & Brdar, 2018). With respect to the survey mode, the SWLS was found invariant across the interviewer-administered and self-administered modes on a German sample (Zhang et al., 2017).

Compared to the SWLS, which is more narrowly focused on the cognitive assessment of one's life in general (Diener et al., 1985), the 14-item MHC-SF represents a wide-encompassing measure of mental health composed of emotional, social, and psychological well-being (Keyes, 2002). While the emotional well-being items measure the frequency of positive affect and thus capture hedonic well-being, the social and psychological well-being items measure individuals' positive functioning in social and private spheres of life and thus capture eudaimonic well-being, which is defined in terms of living a fully functioning life, cultivating personal strengths, and contributing to the common good (Ryan & Deci, 2001). The MHC-SF is widely used in well-being research where, besides the subscale scores, researchers often report the total score representing a measure of general well-being. Several studies examined the psychometric characteristics of the MHC-SF. The majority of

studies confirmed that the three-factor structure is superior to the alternative one- or two-factor structure (e.g., Franken et al., 2018; Joshanloo, 2019; Joshanloo et al., 2017; Petrillo et al., 2015), but several recent studies argued for an even better fit of the bifactor model (e.g., Echeverria et al., 2017; Jovanović, 2015; Lamborn et al., 2018; Monteiro et al., 2020; Žemojtel-Piotrowska et al., 2018), in which general factor is measured by all items, whereas group factors of emotional, social, and psychological well-being account for the residual variance shared by subsets of items. Measurement invariance of the MHC-SF was examined across several demographic groups (e.g., Joshanloo, 2019; Lamborn et al., 2018; Žemojtel-Piotrowska et al., 2018). To our knowledge, no previous study investigated the measurement invariance of the MHC-SF across different survey modes.

Research problem

Evidence suggests that the comparability of results obtained by different modes of data collection may depend both on measurement invariance across modes of administration and characteristics of the construct being measured, such as high sensitivity to social evaluation (De Leeuw & Hox, 2011; Tourangeau et al., 2013), which may lead to a lack of structural invariance. Despite the increasing popularity of mixed-mode approaches to data collection, studies examining measurement invariance of socially sensitive measures of psychological functioning between the traditional interviewer-mediated face-to-face mode and the increasingly popular self-administered web mode in large representative samples of the general population remain sparse. Such samples increase statistical power and provide results that are well generalizable because they reflect the characteristics of the entire population. An exception in this area is the aforementioned study by Zhang et al. (2017). Although their study is commendable in many respects, the results of the cross-mode measurement invariance tests for various mental health and mental illness scales were only briefly described, and the exact statistics were not reported. Strict and structural invariance were not tested, and when comparing survey modes in search of socially desirable response behavior, the focus was on manifest variables. In addition, the sampling frame was different for the different survey modes (e.g., all available telephone numbers in Germany, representatively selected households in Germany).

To advance knowledge in mixed-mode approaches to surveys of socially sensitive constructs of mental health and mental illness and to overcome the shortcomings of previous research in this area, the present study had two main objectives. First, measurement invariance across face-to-face and web modes was examined for the three scales measuring depression (PHQ-8), satisfaction with life (SWLS), and mental health/well-being (MHC-SF) using data from a large

representative sample of the Slovenian population collected as part of the EHIS 3, with the population register serving as the sampling frame. Second, structural invariance was examined by comparing the latent means for the three constructs between the face-to-face and web modes to search for potential systematic shifts in response distribution due to social desirability bias. Based on previous studies that obtained measures of socially sensitive and conceptually related constructs through face-to-face and web modes (e.g., Zhang et al., 2017; see also Heerwegh & Loosveldt, 2011; Klausch et al., 2013; Revilla, 2013), we expected that cross-mode measurement invariance would be established for all the three measures of interest. However, given the highly sensitive nature of the constructs studied, we expected that higher satisfaction with life and better mental health would be reported in the face-to-face mode compared to the web mode and that more depressive symptoms would be reported in the web mode compared to the face-to-face mode. Additionally, we examined the convergent validity of the PHQ-8, SWLS, and MHC-SF based on their intercorrelations, and tested whether the correlations between the three scales were equal in both modes. Finally, our study also provides a novel methodological insight by suggesting a simple transformation that researchers can use to bring data collected by different survey modes onto the same scale when measurement invariance is detected, but significant latent mean differences are found.

Method

Participants and procedure

This study uses data from a large representative sample of the Slovenian population collected in the 2019 European Health Interview Survey (EHIS, Wave 3), conducted on the basis of ‘Commission implementing decision (EU

2018/257’ of 19 February 2018 by Slovenian National Institute of Public Health, which is an authorized producer of national health and healthcare statistics. While the PHQ-8 was part of the statutory questionnaire, the SWLS and MHC-SF were included at the national level. A stratified two-stage probability sample of 16,000 individuals from the general population aged 15 years or older was selected from Central Population Registry that included only individuals living in private households and excluded individuals living in institutional settings such as nursing homes, prisons, convents, etc. Data were collected between March 2019 and November 2019 using a combination of sequential and concurrent mixed-mode designs. 8,000 people were invited to participate in the spring part of the data collection, and a further 8,000 people were invited to participate in the autumn. All individuals selected in the sample received a notification letter inviting them to complete the survey over the web. Individuals who did not respond were contacted by an interviewer for data collection in their homes using computer-assisted face-to-face interviews. The web survey was available online throughout the data collection period. Before consenting to participate, participants were informed about the purpose and procedure of the survey and assured that their responses would be confidential and carefully protected.

In the face-to-face interviews, the interviewer read the questions and possible answers to the respondents. Showcards with answer categories were used to facilitate the response process. In the entry screen, the interviewer could check “I do not know/I do not want to answer” for each question if the respondent specifically said so to the interviewer. There were no mandatory responses in the web survey. If the respondent did not answer a question, the entry “No answer” was noted there. In both cases, the value -1 was stored in the database and treated as a missing value in the data analysis.

A total of 9,900 surveys were conducted with selected individuals; of these, 4,868 were conducted over the web

Table 1 Demographic characteristics of the total sample and separately by face-to-face and web survey modes

	Total sample		F2F		WEB	
	UnW No.	W % / M (SD)	UnW No.	W % / M (SD)	UnW No.	W % / M (SD)
Age						
Mean	9,900	49.2 (18.9)	5,032	55.2 (19.0)	4,868	42.8 (16.4)
Sex						
Male	4,479	49.6%	2,399	50.8%	2,080	48.40%
Female	5,421	50.4%	2,633	49.2%	2,788	51.60%
Education						
Low	1,737	23.0%	1,189	31.0%	548	14.40%
Medium	5,257	52.8%	3,006	56.2%	2,251	49.10%
Higher	2,906	24.2%	837	12.8%	2,069	36.50%

Note. F2F = face-to-face survey mode; WEB = web survey mode. UnW = unweighted; W = weighted

and 5,032 through face-to-face interviews (Table 1). The survey response rate was 67%. Weights were calculated to compensate for differences in the survey response. The final weight of each respondent is the product of non-response weights and calibration weights to adjust for known population values by sex, 10-year age categories with the highest category 80+ years, three levels of education, and statistical region (NUTS level 3). Population data as of 1 January 2019, were used for calibration. In the weighted data, there were between 2.3% and 2.7% missing values for the PHQ-8 items, between 4.6% and 5.4% missing values for the SWLS items, and between 3.1% and 6.6% missing values for the MHC-SF items. Missing values were due to missing responses to certain items for which missing at random (MAR) was assumed, as we see no substantive reasons to expect NMAR missing data; there were also no noticeable missing data patterns. Some missing values were also due to the participation of proxy respondents who were not offered answering items on psychological functioning, but these cases were not included in our analyses.

Table 1 presents the demographic characteristics of the sample. Individuals who participated face-to-face were older than those who participated over the web, $t(9,853.9) = -34.94, p < .001$ (the assumption of the equality of variances was violated, so degrees of freedom were adjusted to correct for unequal variances). In addition, the proportion of male participants was slightly higher in the face-to-face mode than in the web mode, while the proportion of female participants was slightly greater in the web mode than in the face-to-face mode, although the difference was not large, $\chi^2(1) = 6.06, p = .014$. Participants with low (elementary school or less) and medium (secondary school) education were more likely to participate in the face-to-face mode than over the web. Conversely, participants with higher education (tertiary education or more) were more likely to participate over the web than face-to-face, $\chi^2(2) = 895.81, p < .001$.

Measurement instruments

The Patient Health Questionnaire Depression Scale (PHQ-8; Kroenke & Spitzer, 2002; Kroenke et al., 2009, 2010) measures the severity of depressive symptoms and includes eight of the nine criteria on which the diagnosis of depressive disorders is based according to the DSM-5 (American Psychiatric Association, 2013). Respondents indicate, on a four-point scale ranging from ‘never’ (0) to ‘nearly every day’ (3), how often they have experienced each of the eight depressive symptoms described in the items in the past 2 weeks: anhedonia, depressed mood, sleep disturbance, fatigue, appetite changes, low self-esteem, concentration problems, and psychomotor disturbances. An example item is “feeling down, depressed, or hopeless?” Higher scores indicate higher levels of depressive symptoms. A recent

study reported a Cronbach’s alpha of 0.88 for the PHQ-8 (Shin et al., 2019). In a sample of Swedish individuals with systemic sclerosis (Mattsson et al., 2020), the intraclass correlation coefficient (ICC) for the total score was .83, confirming sufficient test-retest reliability over 11 days. A score of ≥ 10 has a sensitivity of 88% and a specificity of 88% for major depression and typically represents clinically significant depression (Kroenke & Spitzer, 2002).

The Satisfaction with Life Scale (SWLS; Diener et al., 1985) measures participants’ cognitive appraisal of their lives in general. The scale contains five items to which participants indicate their agreement on a seven-point scale ranging from “strongly disagree” (1) to “strongly agree” (7). An example item is “In most ways, my life is close to my ideal.” Higher scores indicate higher satisfaction with one’s life. The SWLS has been shown to have high internal consistency with a Cronbach’s alpha of .87 and excellent test-retest reliability with a correlation of .82 over a 2-month period (Diener et al., 1985).

The Mental Health Continuum Short Form (MHC-SF; Keyes, 2002) consists of 14 items that measure three facets of well-being. Three items measure emotional well-being (e.g., “...happy”), five items measure social well-being (e.g., “...that you had something important to contribute to society”), and six items measure psychological well-being (e.g., “...that you had warm and trusting relationships with others”). Respondents rate the frequency of each feeling in the past month on a 6-point scale ranging from ‘never’ (0) to ‘every day’ (5). Higher scores on the subscales and total score indicate higher well-being. A study examining the psychometric properties of MHC-SF (Lamers et al., 2011) yielded the following Cronbach’s alpha coefficients: .83 for emotional and psychological well-being, .74 for social well-being, and .89 for the total score. Crossover path analysis was used to examine the stability of the results. Each subscale consistently predicted that subscale at the 3-month follow-up, with the largest magnitude being approximately .50.

Data analysis

Preliminary steps of data analysis included reviewing the descriptive statistics of the items and scale scores by survey mode, including their means, standard deviations, skewness, and kurtosis. Corrected item-total correlations and Cronbach’s alpha reliability coefficients (α) were calculated for each scale.

Confirmatory factor analyses (CFA) and multiple-group confirmatory factor analyses (MG-CFA) were conducted in *Mplus* version 8.6 (Muthén & Muthén, 1998-2017). The factor structure of the PHQ-8, SWLS, and MHC-SF was first examined separately for the face-to-face and web survey modes. For the PHQ-8 and the SWLS, a theoretically assumed single-factor model was fitted to the data. For the

MHC-SF, two models were compared based on theory and previous research: (a) a model with three correlated factors representing emotional, social, and psychological well-being, which is statistically equivalent to a hierarchical model with a second-order factor accounting for correlations among the three first-order factors and representing general well-being, and (b) a bifactor model with a general well-being factor measured by all items and three uncorrelated group factors for emotional, social, and psychological well-being accounting for residual variance shared by subsets of the items. Mardia's (1970) tests of multivariate skewness and kurtosis were performed for the three scales using the "psych" package in R (Revelle, 2020). All tests were significant, so the robust maximum likelihood estimator (MLR) was used to estimate the models, which is robust in estimating non-normally distributed data and can also handle missing data using full information maximum likelihood (FIML), assuming data are missing at random (MAR; Muthén & Asparouhov, 2002). To achieve model identification, factor variances were constrained to 1 so that all item loadings could be estimated.

To assess the reliability of the factors, McDonald's omega coefficients (ω) were calculated, reflecting the proportion of common variance in the total item variance and corresponding to internal consistency (McDonald, 1999). For the bifactor MHC-SF model, omega-hierarchical (ω_h) and omega-subscale coefficients (ω_s) were also calculated to separate the effects of general and group factors. While ω_h reflects the proportion of total item variance explained by the general factor, ω_s reflects the unique proportion of total item variance accounted for by each group factor controlling for the contribution of the general factor (Reise, 2012). The explained common variance coefficient (ECV) was also calculated for the bifactor MHC-SF model, which is defined as the ratio of the variance explained by the general factor to the total common variance, thus reflecting a relative strength of the general factor to the group factors (Reise et al., 2013).

The best-fitting measurement model of each scale was tested for measurement invariance across face-to-face and web modes. For each scale, the configural invariance model was tested first, which assumes the same factor structure across groups while allowing factor loadings and item intercepts to vary. Second, the metric invariance model was tested, which assumes the same factor loadings across groups, while item intercepts may still vary. Third, the scalar invariance model was tested, which assumes equal factor loadings and item intercepts across groups. Finally, strict invariance was tested, in which equal residual variances are additionally assumed across groups. Achieving scalar invariance is a sufficient condition for comparing latent means between groups (Meredith & Teresi, 2006).

The propensity score adjustment method was used to counterbalance possible confounding effects of

self-selection to survey mode. This method is particularly suitable for reducing the complexity of models with a large number of covariates (Guo & Fraser, 2010; Hox et al., 2015). Because participants were free to choose how they would respond to the survey, different modes could attract participants with different characteristics. Indeed, preliminary inspection of the responses obtained in this study via the face-to-face and web modes revealed an unequal distribution of participants in terms of their sex, age, and education, as reported above (Table 1). In addition, participants' employment status, statistical region, settlement type (varying in size), and period of participation (spring vs. autumn) were used as sociodemographic predictors in a logistic regression analysis to obtain regression-based predicted probabilities of group membership (i.e., propensity scores), which were then used as a single observed covariate in the MG-CFA estimation to control for the sociodemographic composition of the sample. Specifically, all items were regressed on propensity scores, with equality restrictions placed on the regression coefficients across the two survey modes (for a detailed description of the procedure, see Hox et al., 2015). Due to a small number of missing values on employment status (.5%), and thus on the propensity score variable, the number of participants on whom measurement invariance was tested for each scale is slightly smaller than the sum of participants in each survey mode sample included in the single group CFA analyses.

The absolute fit of the CFA and MG-CFA models was assessed using the Satorra-Bentler scaled chi-square test statistic ($SB\chi^2$), root mean square error of approximation (RMSEA) with 90% confidence interval (CI), and standardized root mean square residual (SRMR). According to interpretive guidelines, RMSEA values less than .06 and SRMR values less than .08 were considered indicative of good model fit (Hu & Bentler, 1999), while RMSEA values less than .08 were considered indicative of adequate model fit (Browne & Cudeck, 1992). $SB\chi^2$ values were interpreted with caution, as for very large sample sizes they can be significant even with minor model misspecifications (Cheung & Rensvold, 2002). The incremental fit of the models was assessed using the comparative fit index (CFI), with values greater than .95 indicating good model fit (Hu & Bentler, 1999). We used the Akaike information criterion (AIC) to compare non-nested CFA models, with lower values indicating a better fit to the data than higher values (Akaike, 1973). Because with large sample sizes, χ^2 difference tests can detect even a trivial deterioration in fit, we based our assessment of invariance primarily on the change in fit indices. A difference in fit between successive models with increasingly stringent constraints was assessed with $\Delta RMSEA \leq .015$, $\Delta CFI \leq -.010$, and $\Delta SRMR \leq .030$, as suggested by Chen (2007).

In testing of structural invariance, the web mode was chosen as the reference mode, and its latent means were constrained to 0, whereas they were freely estimated for the face-to-face mode. The freely estimated latent means in the face-to-face mode represent a direct estimate of the difference from the web mode, expressed in units of standard deviation. The significance of the latent mean differences between the two survey modes was assessed using the Wald test.

Finally, to examine the convergent validity of the PHQ-8, SWLS, and MHC-SF, Pearson correlation coefficients between scale scores were calculated separately for the two survey modes. Differences in the magnitude of correlations between the two survey modes were compared using Cohen’s *q* effect sizes. Following Cohen’s (1988) recommendations, values of .10, .30, and .50 were interpreted as small, medium, and large effect sizes, respectively.

Results

Descriptive statistics

First, we examined the descriptive statistics for the three scales and their respective items by mode of data collection (Tables S1-S3 in the Online Supplementary Materials (OSM)). In general, the mean scores of the items measuring satisfaction with life (SWLS) and mental health (MHC-SF) were higher in the face-to-face mode than in the web mode, as were the SWLS total score and the MHC-SF total and subscale scores. The opposite pattern was observed for the mean scores of the items measuring depressive symptoms

(PHQ-8) and for the PHQ-8 total score. The skewness and kurtosis values indicate that the distributions of both item and scale scores generally departed notably from the normal distribution. In some cases, most notably for the MHC-SF scale scores, the shape of the distribution differed across the two survey modes.

The corrected item-total correlations were above .64 for the SWLS items in both modes, with alpha reliability coefficients of .89 and .90 in the face-to-face and web modes, respectively. The emotional well-being subscale items of the MHC-SF had corrected item-total correlations above .81 in both modes, the social well-being subscale items above .58, and the psychological well-being subscale items above .73. The alpha reliability coefficients for emotional well-being were .92/.91, for social well-being .85/.86, and for psychological well-being .92/.92 in the face-to-face and web modes, respectively. For the total score of the MHC-SF, the alpha reliability coefficient was .93 in both modes. The corrected item-total correlations for the PHQ-8 were above .52 in both modes, and the alpha coefficient was .88 in the face-to-face mode and .86 in the web mode.

Single group confirmatory factor analyses

To establish well-fitting baseline models, the fit of the theoretically assumed factor models of the PHQ-8, SWLS, and MHC-SF were first tested separately for the face-to-face and web modes (Table 2). The originally hypothesized single-factor model of the PHQ-8 showed a good fit to the data according to absolute and incremental fit indices for both modes, with the exception of a significant $SB\chi^2$, which was likely due to large sample sizes in both groups. The fit was

Table 2 Confirmatory factor analysis fit statistics for the PHQ-8, SWLS, and MHC-SF by face-to-face and web survey modes

	<i>N</i>	$SB\chi^2_{(df)}$	<i>p</i>	RMSEA	90% CI	CFI	SRMR	AIC
Patient Health Questionnaire Depression Scale								
F2F - single factor	4,900	355.859 ₍₂₀₎	< .001	.059	[.053, .064]	.946	.036	65436.252
WEB - single factor	4,845	280.102 ₍₂₀₎	< .001	.052	[.047, .057]	.957	.033	63945.992
Satisfaction with Life Scale								
F2F - single factor	4,875	366.212 ₍₅₎	< .001	.122	[.111, .132]	.947	.040	78680.868
F2F - single factor (modified)	4,875	36.262 ₍₄₎	< .001	.041	[.029, .053]	.995	.011	78084.191
WEB - single factor	4,673	264.423 ₍₅₎	< .001	.105	[.095, .116]	.957	.032	73563.406
WEB - single factor (modified)	4,673	77.083 ₍₄₎	< .001	.063	[.051, .075]	.988	.017	73137.321
Mental Health Continuum – Short Form								
F2F - three factors/second order	4,860	977.662 ₍₇₄₎	< .001	.050	[.047, .053]	.953	.045	180311.503
F2F - bifactor	4,860	608.472 ₍₆₃₎	< .001	.042	[.039, .045]	.972	.024	179535.052
WEB - three factors/second order	4,849	1381.495 ₍₇₄₎	< .001	.060	[.058, .063]	.947	.044	192869.689
WEB - bifactor	4,849	949.062 ₍₆₃₎	< .001	.054	[.051, .057]	.964	.030	192028.977

Note. F2F = face-to-face survey mode; WEB = web survey mode; $SB\chi^2$ = Satorra-Bentler scaled chi-square; *df* = degrees of freedom; RMSEA = root mean square error of approximation; CI = confidence interval; CFI = comparative fit index; SRMR = standardized root mean square residual; AIC = Akaike information criterion

slightly better for the web mode than for the face-to-face mode, where the upper bound of the RMSEA CI was slightly above the .060 cut-off for good model fit, and the CFI was slightly below the .95 cut-off. Standardized factor loadings for the single-factor solution for both survey modes are presented in Table S4 in the OSM. Factor loadings ranged from .55/.54 to .80/.80 for the face-to-face and web modes, respectively. McDonald's ω reliability coefficients were high, reaching .88 in the face-to-face mode and .86 in the web mode.

The SWLS also has a theoretically assumed single-factor structure, but the single-factor model showed an unsatisfactory fit to the data for both modes, with a large RMSEA value in addition to a significant $SB\chi^2$. The modification indices showed that for both modes, the fit of the model could be improved by releasing the error covariance between items 4 and 5, both of which relate to the evaluation of one's past life. With this error covariance freely estimated, the single-factor model fitted well in both modes, apart from the significant $SB\chi^2$. The fit was slightly better for the face-to-face mode than for the web mode, where the RMSEA was slightly above the cut-off for the good model fit of .060, although the RMSEA CI remained below .080, indicating a satisfactory fit of the model. Table S5 in the OSM shows the standardized factor loadings for the single-factor SWLS model with one freely estimated error covariance. The factor loadings ranged from .60/.70 to .90/.86 for the face-to-face and web modes, respectively. The McDonald's ω reliability coefficient was .88 in the face-to-face mode and .90 in the web mode.

Two competing factor models were tested for the MHC-SF, namely, a model with three correlated factors (mathematically equivalent to the hierarchical model with one second-order factor) and a bifactor model. The fit of the three-factor model was good, but the bifactor model showed an even better fit to the data, with the smallest AIC values and all fit indices except $SB\chi^2$ indicating a good fit. The fit was slightly better for the face-to-face compared to the web mode. Standardized factor loadings for the bifactor model with general factor and the three group factors are presented in Table S6 in the OSM. Factor loadings on the general factor ranged from .59 to .72 for the face-to-face mode and from .56 to .83 for the web mode. For emotional well-being, factor loadings ranged from .54 to .66 for the face-to-face mode and from .57 to .68 for the web mode. The factor loadings on social and psychological well-being were more variable in magnitude. For social well-being, they ranged from .12/.26 to .62/.62, and for psychological well-being, they ranged from .30/-.03 to .49/.52 for the face-to-face and web modes, respectively. The McDonald's ω reliability coefficient for general well-being was .95 for both modes, and for the subscales, it ranged from .86/.86 to .92/.93 for the face-to-face and web modes, respectively. The ECV index

showed a high degree of unidimensionality with the contribution of the general factor to the observed score variance of .67 for the face-to-face mode and .70 for the web mode. ω_h reliability coefficient was .83 for the face-to-face mode and .85 for the web mode, indicating a substantial contribution of the general factor to the total score. On the other hand, the ω_s reliability coefficients were smaller: .43/.46 for emotional well-being, .23/.34 for social well-being, and .24/.03 for psychological well-being for the face-to-face and web modes, respectively. These results suggest that the social and especially psychological well-being subscales have a relatively small share of reliable variance beyond that due to the general factor.

Measurement invariance across face-to-face and web modes

The well-fitting baseline models were combined into MG-CFA models to test the measurement invariance of the three scales studied (Table 3). For the PHQ-8, the measurement invariance of the single-factor model was tested. The configural invariance model, which postulates an invariant factor structure across groups, showed a good fit to the data, as indicated by the RMSEA and SRMR indices, while the CFI was very close to the .95 cut-off for good model fit. As expected, the $SB\chi^2$ was significant, as is common with large sample sizes. The metric invariance model, which postulates invariant factor loadings across groups, fitted the data equally well, with deterioration in model fit compared to the configural model within the change criteria recommended by Chen (2007). The scalar invariance model, which additionally postulates invariant item intercepts across groups, also showed a good fit to the data and no notable deterioration in model fit compared to the metric invariance model. Finally, the strict invariance model with additional constraints on the residual variances of the items fitted the data comparably well as the scalar invariance model.

For the SWLS, the model in which the residual covariance between items 4 and 5 was freely estimated was taken as the baseline model for measurement invariance investigation. The configural invariance model had a good fit to the data, as indicated by all fit indices except the significant $SB\chi^2$. The more restrictive metric invariance model also provided a good fit to the data, and the deterioration in model fit compared to the configural invariance model was small. The fit was also good for the even more restrictive scalar invariance model and the changes in the fit indices compared to the metric invariance model were again within the acceptable range. The same applies to the strict invariance model, where the residual variances and the residual covariance between items 4 and 5 were additionally constrained between groups, compared to the scalar invariance model.

Table 3 Comparison of configural, metric, scalar, and strict invariance models of the PHQ-8, SWLS, and MHC-SF by face-to-face and web survey modes

	N	Goodness-of-fit										Model comparison					
		SB $\chi^2_{(df)}$	p	RMSEA	90% CI	CFI	SRMR	RM	Δ SB $\chi^2_{(df)}$	p	Δ RMSEA	Δ CFI	Δ SRMR				
Patient Health Questionnaire Depression Scale																	
M1p: Configural	9,697	764.167 ₍₄₈₎	< .001	.055	[.052, .059]	.949	.045										
M2p: Metric	9,697	774.798 ₍₅₅₎	< .001	.052	[.049, .055]	.949	.050	26.125 ₍₇₎	< .001	-.003	-.000	.005					
M3p: Scalar	9,697	880.826 ₍₆₂₎	< .001	.052	[.049, .055]	.942	.051	113.961 ₍₇₎	< .001	.000	-.007	.001					
M4p: Strict	9,697	952.733 ₍₇₀₎	< .001	.051	[.048, .054]	.937	.068	87.510 ₍₈₎	< .001	-.001	-.005	.017					
Satisfaction with Life Scale																	
M1s: Configural	9,502	150.916 ₍₁₃₎	< .001	.047	[.041, .054]	.991	.020										
M2s: Metric	9,502	215.030 ₍₁₇₎	< .001	.050	[.044, .056]	.987	.035	65.843 ₍₄₎	< .001	.003	-.004	.015					
M3s: Scalar	9,502	237.725 ₍₂₁₎	< .001	.047	[.041, .052]	.986	.036	13.775 ₍₄₎	.008	-.003	-.001	.001					
M4s: Strict	9,502	252.889 ₍₂₇₎	< .001	.042	[.037, .047]	.986	.031	28.867 ₍₆₎	< .001	-.005	.000	-.005					
Mental Health Continuum – Short Form																	
M1m: Configural	9,661	1614.535 ₍₁₄₀₎	< .001	.047	[.045, .049]	.968	.030										
M2m: Metric	9,661	1916.679 ₍₁₆₄₎	< .001	.047	[.045, .049]	.962	.045	307.739 ₍₂₄₎	< .001	.000	-.006	.015					
M3m: Scalar	9,661	2098.418 ₍₁₇₄₎	< .001	.048	[.046, .050]	.959	.047	190.709 ₍₁₀₎	< .001	.001	-.003	.002					
M4m: Strict	9,661	2629.278 ₍₁₈₈₎	< .001	.052	[.050, .054]	.948	.065	393.134 ₍₁₄₎	< .001	.004	-.011	.018					

Note. RM = reference model; Δ SB χ^2 , Δ RMSEA, Δ CFI, and Δ SRMR = change in fit indices between contiguous nested models; SB χ^2 = Satorra–Bentler scaled chi-square test statistic; *df* = degrees of freedom; RMSEA = root mean square error of approximation; CI = confidence interval; CFI = comparative fit index; SRMR = standardized root mean square residual

For the MHC-SF, measurement invariance was tested for the bifactor model with one general factor and three group factors representing emotional, social, and psychological well-being. The results were similar to those for the PHQ-8 and SWLS. A good fit of the configural invariance model was found, although the $SB\chi^2$ was again significant. The fit was also good for the increasingly restrictive metric, scalar, and strict invariance models, and changes in fit indices between contiguous models with increasingly stringent constraints were within the recommended change criteria. An exception was the ΔCFI for the strict invariance model compared with the scalar invariance model, which slightly exceeded the value of .010 recommended by Chen (2007). However, according to the $\Delta RMSEA$ and $\Delta SRMS$, the deterioration in model fit was also acceptable for the strict invariance model.

In summary, strict invariance was fully confirmed for the three scales, indicating that they are psychometrically equivalent (i.e., they have the same factor structure, factor loadings, item intercepts, and item residual variances) regardless of the mode of administration, and that the underlying constructs of depression, satisfaction with life, and mental health/well-being are understood and measured in the same way in the face-to-face and web modes.

Latent mean differences between modes of data collection

Since strict invariance had been established for the three scales examined, differences in the latent means could be compared between the two survey modes. The latent means were constrained to zero in the web mode, and they were freely estimated in the face-to-face mode. Generally, higher well-being and lower ill-being were reported in the face-to-face compared to the web mode. The latent mean for current depression level (PHQ-8) was $-.253 SD$ lower, and the latent mean for satisfaction with life (SWLS) was $.402 SD$ higher in the face-to-face compared to the web mode, with both

differences being statistically significant ($p < .001$). While no significant differences were observed in latent means of emotional and psychological well-being (MHC-SF) between the two modes, the latent mean for social well-being was $.263 SD$ higher, and the latent mean for general well-being was $.605 SD$ higher in the face-to-face compared to the web mode ($p < .001$). These values equal the values of Cohen's d , applied to latent means, as discussed by Hancock (2001).

Convergent validity of the PHQ-8, SWLS, and MHC-SF scale scores

Correlations between the PHQ-8, SWLS, and MHC-SF scale scores were broadly similar across the two survey modes (Table 4), with very small Cohen's q effect sizes for the differences in correlations between the two modes ranging from $-.04$ to $.07$. The PHQ-8, which measures the severity of current depressive symptoms, had a medium negative correlation with satisfaction with life as measured by the SWLS and a medium to large correlation with general well-being as measured by the MHC-SF total score. The PHQ-8 also had a large negative correlation with the emotional well-being subscale of the MHC-SF and medium correlations with the social and psychological well-being subscales. The SWLS had a large correlation with general well-being and the emotional well-being subscale of the MHC-SF and medium to large correlations with the social and psychological well-being subscales of the MHC-SF. The correlations between the MHC-SF subscales and the total score were all large.

Discussion

In this study, we sought to investigate mode inherent effects that can be observed when a combination of different survey modes is used to measure socially sensitive constructs. Specifically, using a large representative sample of the Slovenian population, we examined whether depressive symptoms,

Table 4 Pearson correlations between the PHQ-8, SWLS, and MHC-SF by face-to-face and web survey modes

	PHQ-8	SWLS	MHC-SF Emotional	MHC-SF Social	MHC-SF Psychological	MHC-SF General
PHQ-8	/	-.43	-.54	-.32	-.42	-.46
SWLS	-.46	/	.56	.46	.44	.55
MHC-SF Emotional	-.59	.62	/	.55	.58	.77
MHC-SF Social	-.37	.45	.54	/	.65	.89
MHC-SF Psychological	-.48	.51	.64	.68	/	.89
MHC-SF General	-.53	.59	.77	.89	.92	/

Note. Correlations for the face-to-face mode above the diagonal, correlations for the web mode below the diagonal. All correlations are significant at $p < .001$

satisfaction with life, and mental health/well-being, as assessed by the PHQ-8, SWLS, and MHC-SF, respectively, are measured in the same way in face-to-face and web modes after controlling for respondent differences in sociodemographic variables. In addition, we examined whether latent mean differences exist between the two modes that can most likely be attributed to more socially desirable responses in the presence of an interviewer in a person-to-person communication compared to responses over the web. Our results confirmed our expectations: while strict measurement invariance was established for all three scales across the two survey modes, statistically significant and notable differences in latent means were observed, suggesting that individuals who respond face-to-face systematically report better psychological functioning compared to those who respond over the web.

Factor structure of the scales is consistent with previous research

The first aim of this study was to examine the measurement invariance of the PHQ-8, SWLS, and MHC-SF across face-to-face and web survey modes. However, before measurement invariance could be examined, good-fitting baseline models had to be established for each scale. For the PHQ-8, a theoretically proposed single-factor model showed a good fit to the data in agreement with several previous studies (Baas et al., 2011; Galenkamp et al., 2017; Keum et al., 2018; López-Torres et al., 2019). Although in some studies various versions of the two-factor model showed an even better fit to the data, the correlations between factors were usually very high, indicating poor discrimination between factors (Boothroyd et al., 2019; Keum et al., 2018; Lamela et al., 2020; Pagán-Torres et al., 2020). Consistent with theoretical expectations and previous work (Emerson et al., 2017; Hinz et al., 2018; Jang et al., 2017; Jovanović & Brdar, 2018), a single-factor model also provided a good fit to the SWLS data, but only after relaxing a residual covariance between items 4 and 5. The same residual covariance was also freely estimated in some previous studies to achieve a satisfactory model fit (Bai et al., 2011; Clench-Aas et al., 2011; Jovanović, 2019). This decision was based on the reasoning that both items tap individuals' assessment of their past accomplishments, whereas the first three items assess present life satisfaction. Finally, for the MHC-SF, a better fit was obtained for the bifactor model, thus supporting findings from several recent studies on the factor structure of the MHC-SF (Echeverría et al., 2017; Jovanović, 2015; Lamborn et al., 2018; Monteiro et al., 2020; Žemojtel-Piotrowska et al., 2018). However, it should be noted that the group factors were relatively weak compared to the general factor, with over two-thirds of the explained item variance due to the general factor and only less than one-third of the

variance due to emotional, social, or psychological well-being. In addition, the subscale scores added little unique, reliable variance beyond the general factor, suggesting limited usefulness of the subscale scores when formed together with the total score, which has already been observed by some other authors (e.g., Jovanović, 2015; Monteiro et al., 2020; Žemojtel-Piotrowska et al., 2018).

Measurement invariance confirms metric equivalence across face-to-face and web modes

In the next step, measurement invariance was tested for the three scales across face-to-face and web survey modes to identify possible mode inherent effects in the question-answer process. Since mode effects can also occur because different types of participants are more likely to respond in different modes (e.g., because they do not have access to the internet or are unfamiliar with technology), we controlled for the self-selection effects using propensity scores based on multiple demographic control variables (Hox et al., 2015). Multiple-group confirmatory factor analyses (MG-CFAs) were conducted using the previously established good-fitting baseline factor models of the three scales studied. All three scales achieved strict invariance across face-to-face and web modes, suggesting that the question-answer process was similar in both survey modes. In other words, participants understood and responded to the items measuring depressive symptoms (PHQ-8), satisfaction with life (SWLS), and mental health/well-being (MHC-SF) in a similar manner regardless of the mode of data collection. The use of show cards in face-to-face mode could represent an important aspect of similarity to the web mode, potentially contributing to measurement invariance between the two modes.

These findings comply with the results of a study in a representative German sample (Zhang et al., 2017), which confirmed full or partial scalar measurement invariance for 13 different scales measuring positive and negative aspects of mental health across face-to-face, telephone, web, and paper-and-pencil modes. The study also included the SWLS for which full scalar invariance was found. Whereas no previous study examined cross-mode measurement invariance of the MHC-SF, scalar measurement invariance across telephone and face-to-face modes was found for the PHQ-9 in a large sample of British patients enrolled in psychological therapies for depression and anxiety disorders (Ryan et al., 2013). Our findings from a large representative Slovenian general population sample add to this sparse previous research on cross-mode measurement invariance of scales measuring various aspects of psychological functioning and mental health by showing that the PHQ-8, SWLS, and MHC-SF are metrically equivalent when completed face-to-face and over the web. Earlier research also found full scalar invariance across face-to-face and web modes for measures

of trust and attitudes toward immigration in representative Dutch samples (Revilla, 2013). However, another study using representative Dutch data did not confirm invariance of attitudinal scales across interviewer-administered (face-to-face and telephone) and self-completion modes (web and paper-and-pencil; Klausch et al., 2013).

Our findings are important in light of the increasing use of more than one mode of data collection in an attempt to improve coverage and recruit as many participants as possible (De Leeuw, 2018; De Leeuw et al., 2019; Dillman et al., 2014; Tourangeau, 2017; Tourangeau et al., 2013). In particular, different modes of data collection may have different effects on responses, thus compromising the validity of cross-mode comparisons and calling into question the practice of pooling data across modes (De Leeuw et al., 2019; Hox et al., 2015; Tourangeau, 2017; Tourangeau et al., 2013). Our results demonstrated that responses on the PHQ-8, SWLS, and MHC-SF obtained through face-to-face and web modes are metrically equivalent, which is a necessary prerequisite for comparing scale scores and their relationships with other relevant constructs across these modes. However, researchers are often interested in pooling data collected using different modes, but before this can be done, differences in latent means between modes must also be examined.

Comparison of mean levels indicates a presence of social desirability bias in face-to-face mode

After establishing measurement invariance, we were able to compare the latent means across the two survey modes (Meredith & Teresi, 2006). Significant differences were observed after self-selection effects were adjusted for using propensity scores, with participants responding face-to-face reporting significantly lower depression levels, higher satisfaction with life, and higher social and general well-being compared to those responding over the web. This is a prevalent observation in studies examining mode effects (Tourangeau & Yan, 2007), and it is also consistent with results from a comparable study by Zhang et al. (2017), who found significant latent mean differences in the same direction for different scales measuring positive and negative aspects of mental health, including the SWLS. These results cannot be attributed to different sample composition due to self-selection effects, but rather suggest a systematic bias due to greater social desirability in modes involving an interviewer, shifting the distribution of responses to the items tapping socially sensitive constructs in a more socially desirable direction (De Leeuw & Hox, 2011; Tourangeau et al., 2013). Specifically, the presence of an interviewer may bring to one's awareness a potential social evaluation threat resulting in a tendency to adapt answers to socially sensitive questions in a manner that will be viewed favorably by the interviewer

(Dillman et al., 2014; Kreuter et al., 2009; Tourangeau et al., 2013). However, latent means for emotional and psychological well-being did not differ between face-to-face and web modes. We should first note that the group factors were rather weak and contributed little to the explained item variance, whereas the general factor of the MHC-SF items was quite strong. For social well-being, the difference between survey modes was still significant, perhaps because society sets stricter social norms regarding the individual's contribution to the common good, whereas emotional and psychological well-being refer to private aspects, where social pressure seems to be weaker.

The strict measurement invariance achieved in our study for the PHQ-8, SWLS, and MHC-SF also implies that the comparison of composite scores for these scales is justified between face-to-face and web modes. However, the systematic bias toward reporting greater well-being in person-to-person communication than over the web should be taken into account. This bias is manifested mainly as the difference between the latent means but also indirectly as the differences in the shapes of the scale scores distributions. Most notably, the PHQ-8 scale scores distribution exhibited greater right skewness and was more leptokurtic in the face-to-face mode than in the web mode, with the same pattern observed for the distributions of the PHQ-8 items. Nonetheless, internal consistency coefficients were high in both survey modes, indicating high reliability of scale scores obtained both face-to-face and over the web. These results are in contrast to some previous studies that reported higher reliability/internal consistency in the self-administered modes (paper-and-pencil/web) than in the interviewer-administered modes (face-to-face/telephone; e.g., Klausch et al., 2013). This disparity is likely due to the fact that all questions are presented together on a single sheet of paper or screen, which might result in positive error correlations and consequently in artifactually high values of internal consistency.

Despite the absence of measurement non-invariance, our findings on the different mean levels of depressive symptoms, satisfaction with life, and mental health/well-being reported in the interviewer- and self-administered survey modes indicate the presence of social desirability bias and caution against simply pooling data on these socially sensitive constructs collected in face-to-face and web modes. In general, pooling data from both sources would mean that participants who responded in the face-to-face mode would be systematically assessed as having lower depression levels and higher satisfaction with life and general and social well-being than participants in the web mode with the same levels of the latent traits. Pooling should therefore be avoided when individual comparisons are of interest and when univariate statistics are reported, although it may be warranted in correlational analyses when controlling for survey mode.

Convergent validity is stable across modes of application

The additional aim of the present study was to investigate the convergent validity of the PHQ-8, SWLS, and MHC-SF by examining their intercorrelations, and to test whether these correlations were similar regardless of the mode of administration. Very small and practically negligible differences were found in the magnitude of the correlations between the two modes. The correlations were comparable to those found in previous studies (e.g., Keum et al., 2018; Perugini et al., 2017; Yun et al., 2019), suggesting good convergent validity of the three scales for both survey modes. In particular, depressive symptoms showed the strongest negative correlation with emotional well-being, which may be explained by the absence of positive emotions typical of depression (Watson et al., 1988). Satisfaction with life also showed the strongest positive correlation with emotional well-being, which was expected since both scales tap hedonic elements of well-being (Ryan & Deci, 2001). Both depressive symptoms and satisfaction with life also had large correlations in the opposite direction with general well-being, as expected. None of the correlations were too large to question the discriminant validity of the scales.

Implications for mixed-mode research

Our results have important implications in the area of mixed-mode research. First, because strict measurement invariance was found for the PHQ-8, SWLS, and MHC-SF, researchers can validly compare structural relationships between depression, satisfaction with life, and mental health/well-being and other relevant constructs, as well as latent means of these measures, between face-to-face and web modes, preferably after controlling for (self-)selection effects. Direct mean comparisons between the two modes are also possible, but again controlling for differences in sample composition is needed. While bivariate and multivariate statistics can be performed on the pooled data collected with the PHQ-8, SWLS, and MHC-SF across face-to-face and web modes as long as the mode effects are controlled, our results advise against pooling data when reporting univariate statistics because of a non-negligible bias toward socially desirable responses in an interviewer-mediated face-to-face mode. These findings also imply that if norms are to be developed for the three scales, they should differ depending on the mode of data collection.

According to de Leeuw et al. (2019), mode effects in surveys with a combination of interviewer-mediated and self-administered modes when measuring socially sensitive constructs can only be compensated for by including a special section in which all respondents across modes answer sensitive questions in a self-administered format. However, if

researchers are interested in individual comparisons or want to report univariate statistics for data pooled across different modes, they can also use a simple ad hoc transformation to bring the data to the same scale that we suggest in the OSM.

Limitations and future directions

Among the limitations of this study, we should first note that the distinction between measurement and structural invariance was not completely clear in our case. The two groups compared were not pre-existing groups but were rather defined by the data collection mode. One possible viewpoint would be that the average level of the measured trait (for instance, well-being) is equal in both groups, especially after controlling for self-selection propensities. Technically, this would imply constraining both latent trait means to equality and letting all intercepts for one of the groups increase by a constant amount. Therefore, the mode effect could be modeled as the lack of the scalar invariance rather than as the lack of the structural invariance. However, this would obscure the fact that the pattern of the intercepts is approximately equal in both groups and that the mode effect results mainly in the location shift of the latent trait distribution. Therefore, we considered it more informative to model the mode effect as the lack of structural invariance.

Because this was not a controlled experiment, we could not completely rule out the effects of self-selection in different survey modes. Although we used a propensity score method to control for a number of demographic characteristics, other unaccounted demographic variables could be relevant covariates. Indeed, our logistic regression model was able to correctly classify 70.0% of participants into the two survey modes, leaving room for other covariates, not just demographic, to provide additional information and improve classification accuracy (e.g., Hox et al., 2015). Another limitation related to covariate adjustment is a small number of missing values in one of the control variables, which was self-reported (all other control variables were taken from the population register), due to which measurement invariance was not tested on the whole sample. We also had some missing values on the items of the three scales studied, but the MLR procedure used to estimate the models is able to handle this kind of missing data (Muthén & Asparouhov, 2002).

In this study, we hypothesized that social desirability induced by the presence of an interviewer plays a role in the differences we found between face-to-face and web modes of survey administration, but we did not conduct social desirability assessments to confirm this assumption. Future studies of cross-mode measurement invariance of measures of socially sensitive constructs should therefore include such assessments to confirm social desirability bias as a possible explanation for mean differences between modes.

Although our results indicate that the PHQ-8, SWLS, and MHC-SF scale scores can be meaningfully compared between face-to-face and web survey modes, other survey modes, such as telephone and paper-and-pencil, are still commonly used in mixed-mode surveys. Future studies on representative population samples are needed to extend our findings to these modes (e.g., Zhang et al., 2017) and to other scales used to measure various psychological constructs in mixed-mode surveys. Future research would also benefit from studies that thoroughly investigate the reasons for mode-inherent effects, possibly using a qualitative methodology. The role of show cards in the face-to-face survey mode, and their contribution to mode-inherent effects could also be clarified by future studies. Finally, tests of the measurement invariance of the PHQ-8, SWLS, and MHC-SF across different demographic groups and over time would enrich knowledge about the measurement properties of the Slovenian versions of these scales.

Conclusion

In conclusion, the results of this study are in line with previous findings and show significant effects of the administration mode on the level of self-reported depressive symptoms, satisfaction with life, and mental health/well-being as measured by the PHQ-8, SWLS, and MHC-SF, respectively. These mode effects cannot be attributed to measurement non-invariance of the scales across face-to-face and web survey modes, but most likely to greater social desirability of responses to different items in person-to-person communication, which establishes an immediate context of social evaluation that is much less prominent in the web survey mode. Administration mode effects need to be tested and taken into consideration when interpreting results obtained through different survey modes, especially when using measures of culturally desirable traits and behaviors, such as mental health and well-being.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.3758/s13428-022-01867-8>.

Acknowledgements This work was supported by the Slovenian Research Agency (research core funding No. P5-0062 and P5-0110) and Eurostat (grant agreement No. 07154.2017.002-2017.500 for ‘Pre-testing of new variables for future European Health Interview Survey waves’). Appreciation is expressed to Maruša Rehberger for her help with data preparation.

Declarations

Conflicts of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Ethics approval This study uses secondary data from a European Health Interview Survey (EHIS, Wave 3), conducted on the basis of ‘Commission implementing decision (EU) 2018/257’ of 19 February 2018 by Slovenian National Institute of Public Health, which is an authorized producer of national health and health care statistics.

Consent to participate Based on the information about the purpose and procedure of the survey and the assurance that responses would be confidential and carefully protected, individuals consented to participate before answering the survey questions.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N., & Csaki, F. (Eds.), *2nd International Symposium on Information Theory* (pp. 267–281). Akademiai Kiado.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Publishing.
- Baas, K. D., Cramer, A. O., Koeter, M. W., Van De Lisdonk, E. H., Van Weert, H. C., & Schene, A. H. (2011). Measurement invariance with respect to ethnicity of the Patient Health Questionnaire–9 (PHQ–9). *Journal of Affective Disorders*, *129*(1–3), 229–235. <https://doi.org/10.1016/j.jad.2010.08.026>
- Bäckström, M., & Björklund, F. (2014). Social desirability in personality inventories: The nature of the evaluative factor. *Journal of Individual Differences*, *35*(3), 144–157. <https://doi.org/10.1027/1614-0001/a000138>
- Bäckström, M., Björklund, F., & Larsson, M. R. (2012). Social desirability in personality assessment: Outline of a model to explain individual differences. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 201–213). Oxford University Press.
- Bai, X., Wu, C., Zheng, R., & Ren, X. (2011). The psychometric evaluation of the Satisfaction with Life Scale using a nationally representative sample of China. *Journal of Happiness Studies*, *12*(2), 183–197. <https://doi.org/10.1007/s10902-010-9186-x>
- Berzelak, N., & Vehovar, V. (2018). Mode effects on socially desirable responding in web surveys compared to face-to-face and telephone surveys. *Metodoloski zvezki*, *15*(2), 21–43.
- Boothroyd, L., Dagnan, D., & Muncer, S. (2019). PHQ–9: One factor or two? *Psychiatry Research*, *271*, 532–534. <https://doi.org/10.1016/j.psychres.2018.12.048>
- Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258. <https://doi.org/10.1177/0049124192021002005>
- Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466.
- Callegaro, M., Manfreda, K. L., & Vehovar, V. (2015). *Web survey methodology*. Sage.
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504. <https://doi.org/10.1080/10705510701301834>
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5
- Christian, L. M., Dillman, D. A., & Smyth, J. D. (2007). Helping respondents get it right the first time: the influence of words, symbols, and graphics in web surveys. *Public Opinion Quarterly*, *71*(1), 113–125. <https://doi.org/10.1093/poq/nfl039>

- Clench-Aas, J., Nes, R. B., Dalgard, O., & Aarø, L. (2011). Dimensionality and measurement invariance in the Satisfaction with Life Scale in Norway. *Quality of Life Research*, 20(8), 1307–1317. <https://doi.org/10.1007/s11136-011-9859-x>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge Academic.
- De Leeuw, E. D. (1992). *Data quality in mail, telephone and face to face surveys*. TT Publikaties, Plantage Daklaan 40, 1018CN Amsterdam.
- De Leeuw, E. D. (2018). Mixed-Mode: Past, Present, and Future. *Survey Research Methods*, 12(2), 75–89. <https://doi.org/10.18148/srm/2018.v12i2.7402>
- De Leeuw, E. D., & Hox, J. J. (2011). Internet surveys as part of a mixed-mode design. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and Behavioral Research and the Internet* (pp. 45–76). Routledge.
- De Leeuw, E. D., Suzer-Gurtekin, Z. T., & Hox, J. J. (2019). The design and implementation of mixed-mode surveys. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Wiley series in survey methodology. Advances in comparative survey methods: Multicultural, multiregional, and multicultural contexts (3MC)* (pp. 387–408). John Wiley & Sons Inc.
- Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, 49(1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13
- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys*. Wiley.
- Echeverria, G., Torres, M., Pedrals, N., Padilla, O., Rigotti, A., & Bitran, M. (2017). Validation of a Spanish version of the Mental Health Continuum–Short Form Questionnaire. *Psicothema*, 29(1), 96–102. <https://doi.org/10.7334/psicothema>
- Emerson, S. D., Guhn, M., & Gadermann, A. M. (2017). Measurement invariance of the Satisfaction with Life Scale: Reviewing three decades of research. *Quality of Life Research*, 26, 2251–2264. <https://doi.org/10.1007/s11136-017-1552-2>
- Franken, K., Lamers, S. M. A., Ten Klooster, P. M., Bohlmeijer, E. T., & Westerhof, G. J. (2018). Validation of the Mental Health Continuum–Short Form and the dual continua model of well-being and psychopathology in an adult mental health setting. *Journal of Clinical Psychology*, 74(12), 2187–2202. <https://doi.org/10.1002/jclp.22659>
- Galenkamp, H., Stronks, K., Snijder, M. B., & Derks, E. M. (2017). Measurement invariance testing of the PHQ–9 in a multi-ethnic population in Europe: the HELIUS study. *BMC Psychiatry*, 17(1), 1–14. <https://doi.org/10.1186/s12888-017-1506-9>
- Guo, S., & Fraser, M. W. (2010). *Propensity Score Analysis. Statistical Methods and Applications*. SAGE publications.
- Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, 66(3), 373–388. <https://doi.org/10.1007/BF02294440>
- Heerwegh, D., & Loosveldt, G. (2011). Assessing mode effects in a national crime victimization survey using structural equation models: social desirability bias and acquiescence. *Journal of Official Statistics*, 27(1), 49–63.
- Hinz, A., Conrad, I., Schroeter, M. L., Glaesmer, H., Brähler, E., Zenger, M., et al. (2018). Psychometric properties of the Satisfaction with Life Scale (SWLS), derived from a large German community sample. *Quality of Life Research*, 27(6), 1661–1670. <https://doi.org/10.1007/s11136-018-1844-1>
- Hox, J., De Leeuw, E., & Klausch, T. (2017). Mixed mode research: issues in design and analysis. In P. P. Biemer, E. De Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West, *Total survey error in practice* (pp. 511–530). John Wiley & Sons, Inc.
- Hox, J. J., De Leeuw, E. D., & Zijlman, E. A. O. (2015). Measurement equivalence in mixed mode surveys. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00087>
- Hu, L., & Bentler, P. M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the Effect of Data Collection Mode on Measurement. *International Statistical Review*, 78(1), 3–20. <https://doi.org/10.1111/j.1751-5823.2010.00102.x>
- Jang, S., Kim, E. S., Cao, C., Allen, T. D., Cooper, C. L., Lapierre, L. M., et al. (2017). Measurement invariance of the satisfaction with life scale across 26 countries. *Journal of Cross-Cultural Psychology*, 48(4), 560–576. <https://doi.org/10.1177/0022022117697844>
- Joshanloo, M. (2019). Factor structure and measurement invariance of the MHC–SF in the USA. *European Journal of Psychological Assessment*, 35(4), 521–525. <https://doi.org/10.1027/1015-5759/a000425>
- Joshanloo, M., Jose, P. E., & Kiehlkowski, M. (2017). The value of exploratory structural equation modeling in identifying factor overlap in the Mental Health Continuum–Short Form (MHC–SF): A study with a New Zealand sample. *Journal of Happiness Studies*, 18(4), 1061–1074. <https://doi.org/10.1007/s10902-016-9767-4>
- Jovanović, V. (2015). Structural validity of the Mental Health Continuum–Short Form: The bifactor model of emotional, social and psychological well-being. *Personality and Individual Differences*, 75, 154–159. <https://doi.org/10.1016/j.paid.2014.11.026>
- Jovanović, V. (2019). Measurement invariance of the Serbian version of the Satisfaction with Life Scale across age, gender, and time. *European Journal of Psychological Assessment*, 35(4), 555–563. <https://doi.org/10.1027/1015-5759/a000410>
- Jovanović, V., & Brdar, I. (2018). The cross-national measurement invariance of the Satisfaction with Life Scale in a sample of undergraduate students. *Personality and Individual Differences*, 128, 7–9. <https://doi.org/10.1016/j.paid.2018.02.010>
- Keum, B. T., Miller, M. J., & Inkelas, K. K. (2018). Testing the factor structure and measurement invariance of the PHQ–9 across racially diverse US college students. *Psychological Assessment*, 30(8), 1096–1106. <https://doi.org/10.1037/pas0000550>
- Keyes, C. L. M. (2002). The mental health continuum: From languishing to flourishing in life. *Journal of Health and Social Behaviour*, 43(2), 207–222. <https://doi.org/10.2307/3090197>
- Keyes, C. L. M. (2005). Mental illness and/or mental health? Investigating axioms of the complete state model of health. *Journal of Consulting and Clinical Psychology*, 73, 539–548. <https://doi.org/10.1037/0022-006X.73.3.539>
- Keyes, C. L. M. (2007). Promoting and protecting mental health as flourishing: A complementary strategy for improving national mental health. *American Psychologist*, 62(2), 95–108. <https://doi.org/10.1037/0003-066X.62.2.95>
- Klausch, T., Hox, J. J., & Schouten, B. (2013). Measurement Effects of Survey Mode on the Equivalence of Attitudinal Rating Scale Questions. *Sociological Methods & Research*, 42(3), 227–263. <https://doi.org/10.1177/0049124113500480>
- Kreuter, F., Presser, S., & Tourangeau, R. (2009). Social desirability bias in CATI, IVR and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, 72(5), 847–865. <https://doi.org/10.1093/poq/nfn063>
- Kroenke, K., & Spitzer, R. L. (2002). The PHQ–9: A new depression diagnostic and severity measure. *Psychiatric Annals*, 32(9), 509–515. <https://doi.org/10.3928/0048-5713-20020901-06>
- Kroenke, K., Spitzer, R. L., Williams, J. B. W., & Lowe, B. (2010). The Patient Health Questionnaire somatic, anxiety, and depressive symptom scales: A systematic review. *General Hospital Psychiatry*, 32(4), 345–359. <https://doi.org/10.1016/j.genhosppsych.2010.03.006>

- Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1–3), 163–173. <https://doi.org/10.1016/j.jad.2008.06.026>
- Lamborn, P., Cramer, K. M., & Riberdy, A. (2018). The Structural Validity and Measurement Invariance of the Mental Health Continuum – Short Form (MHC-SF) in a Large Canadian Sample. *Journal of Well-Being Assessment*, 2(1), 1–19. <https://doi.org/10.1007/s41543-018-0007-z>
- Lamela, D., Soreira, C., Matos, P., & Morais, A. (2020). Systematic review of the factor structure and measurement invariance of the patient health questionnaire-9 (PHQ-9) and validation of the Portuguese version in community settings. *Journal of Affective Disorders*, 276, 220–233. <https://doi.org/10.1016/j.jad.2020.06.066>
- Lamers, S. M., Westerhof, G. J., Bohlmeijer, E. T., ten Klooster, P. M., & Keyes, C. L. (2011). Evaluating the psychometric properties of the mental health continuum-short form (MHC-SF). *Journal of Clinical Psychology*, 67(1), 99–110. <https://doi.org/10.1002/jclp.20741>
- López-Torres, S., Pérez-Pedrogo, C., Sánchez-Cardona, I., & Sánchez-Cesáreo, M. (2019). Psychometric Properties of the PHQ-A among a Sample of Children and Adolescents in Puerto Rico. *Current Psychology*, 1–9. <https://doi.org/10.1007/s12144-019-00468-7>
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3), 519–530.
- Mattsson, M., Sandqvist, G., Hesselstrand, R., Nordin, A., & Boström, C. (2020). Validity and reliability of the Patient Health Questionnaire-8 in Swedish for individuals with systemic sclerosis. *Rheumatology International*, 40(10), 1675–1687. <https://doi.org/10.1007/s00296-020-04641-1>
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Psychology Press.
- Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical care*, 44(11 Suppl 3), S69–S77. <https://doi.org/10.1097/01.mlr.0000245438.73837.89>
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge/Taylor & Francis Group.
- Monteiro, F., Fonseca, A., Pereira, M., & Canavarro, M. C. (2020). Measuring Positive Mental Health in the Postpartum Period: The Bifactor Structure of the Mental Health Continuum-Short Form in Portuguese Women. *Assessment*. <https://doi.org/10.1177/1073191120910247>
- Muthén, B. O., & Asparouhov, T. (2002). *Using Mplus Monte Carlo simulations in practice: A note on non-normal missing data in latent variable models*. Version 2, March 22, 2002.
- Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.
- OECD (2020). *How's Life? 2020: Measuring Well-being*. Available online at <https://www.oecd-ilibrary.org/sites/9870c393-en/index.html?itemId=/content/publication/9870c393-en> (Accessed June 11, 2020)
- Pagán-Torres, O. M., González-Rivera, J. A., & Rosario-Hernández, E. (2020). Psychometric analysis and factor structure of the Spanish version of the eight-item Patient Health Questionnaire in a general sample of Puerto Rican adults. *Hispanic Journal of Behavioral Sciences*, 42(3), 401–415. <https://doi.org/10.1177/0739986320926524>
- Perinelli, E., & Gremigni, P. (2016). Use of social desirability scales in clinical psychology: A systematic review. *Journal of clinical psychology*, 72(6), 534–551. <https://doi.org/10.1002/jclp.22284>
- Perugini, M. L. L., de la Iglesia, G., Solano, A. C., & Keyes, C. L. M. (2017). The mental health continuum-short form (MHC-SF) in the Argentinean context: Confirmatory factor analysis and measurement invariance. *Europe's Journal of Psychology*, 13(1), 93–108. <https://doi.org/10.5964/ejop.v13i1.1163>
- Petrillo, G., Capone, V., Caso, D., & Keyes, C. L. M. (2015). The Mental Health Continuum-Short Form (MHC-SF) as a measure of well-being in the Italian context. *Social Indicators Research*, 121(1), 291–312. <https://doi.org/10.1007/s11205-014-0629-3>
- Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47, 667–696. <https://doi.org/10.1080/00273171.2012.715555>
- Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modelling a bifactor perspective. *Educational and Psychological Measurement*, 73, 5–26. <https://doi.org/10.1177/0013164412449831>
- Revelle, W. (2020). *psych: Procedures for Personality and Psychological Research*. Northwestern University <https://CRAN.R-project.org/package=psych>.
- Revilla, M. A. (2013). Measurement invariance and quality of composite scores in a face-to-face and a web survey. *Survey Research Methods*, 7(1), 17–28. <https://doi.org/10.18148/srm/2013.v7i1.5098>
- Ryan, R. M., & Deci, E. L. (2001). On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology*, 52, 141–166. <https://doi.org/10.1146/annurev.psych.52.1.141>
- Ryan, T. A., Bailey, A., Fearon, P., & King, J. (2013). Factorial invariance of the patient health questionnaire and generalized anxiety disorder questionnaire. *British Journal of Clinical Psychology*, 52(4), 438–449. <https://doi.org/10.1111/bjc.12028>
- Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist*, 55(1), 5–14. <https://doi.org/10.1037/0003-066X.55.1.5>
- Shin, C., Lee, S. H., Han, K. M., Yoon, H. K., & Han, C. (2019). Comparison of the Usefulness of the PHQ-8 and PHQ-9 for Screening for Major Depressive Disorder: Analysis of Psychiatric Outpatient Data. *Psychiatry Investigation*, 16(4), 300–305. <https://doi.org/10.30773/pi.2019.02.01>
- Spitzer, R. L., Kroenke, K., Williams, J. B., & Patient Health Questionnaire Primary Care Study Group. (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *Jama*, 282(18), 1737–1744. <https://doi.org/10.1001/jama.282.18.1737>
- Suldo, S. M., & Shaffer, E. J. (2008). Looking beyond psychopathology: The dual-factor model of mental health in youth. *School Psychology Review*, 37(1), 52–68. <https://doi.org/10.1080/02796015.2008.12087908>
- Tourangeau, R. (2017). Mixing Modes: Tradeoffs among Coverage, Non-response, and Measurement Error. In P. P. Biemer, E. De Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West (Eds.), *Total Survey Error in Practice* (pp. 115–132). John Wiley & Sons.
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133, 859–883. <https://doi.org/10.1037/0033-2909.133.5.859>
- Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The science of web surveys*. Oxford University Press.
- Veenhoven, R. (2020). *World Database of Happiness*. Erasmus University Rotterdam, The Netherlands. Available online at: <http://worlddatabaseofhappiness.eur.nl> (Accessed June 11, 2020)
- Watson, D., Clark, L. A., & Carey, G. (1988). Positive and negative affectivity and their relation to anxiety and depressive disorders. *Journal of Abnormal Psychology*, 97(3), 346–353. <https://doi.org/10.1037/0021-843X.97.3.346>
- Westerhof, G. J., & Keyes, C. L. (2010). Mental illness and mental health: The two continua model across the lifespan. *Journal of Adult Development*, 17(2), 110–119. <https://doi.org/10.1007/s10804-009-9082-y>

- Yun, Y. H., Rhee, Y. E., Kang, E., & Sim, J. A. (2019). The satisfaction with life scale and the subjective well-being inventory in the general Korean population: psychometric properties and normative data. *International Journal of Environmental Research and Public Health*, *16*(9), 1538. <https://doi.org/10.3390/ijerph16091538>
- Żemojtel-Piotrowska, M., Piotrowski, J. P., Osin, E. N., Ciecuch, J., Adams, B. G., Ardi, R., et al. (2018). The mental health continuum-short form: The structure and application for cross-cultural studies—A 38 nation study. *Journal of Clinical Psychology*, *74*(6), 1034–1052. <https://doi.org/10.1002/jclp.22570>
- Zhang, X., Kuchinke, L., Woud, M. L., Velten, J., & Margraf, J. (2017). Survey method matters: Online/offline questionnaires and

face-to-face or telephone interviews differ. *Computers in Human Behavior*, *71*, 172–180. <https://doi.org/10.1016/j.chb.2017.02.006>

Open practices statement The data that support the findings of this study are available from the Slovenian National Institute of Public Health upon request (<https://www.nijz.si/en>). Neither the study nor the analyses were preregistered.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.